DOCUMENT RESUME

ED 255 566 TM 850 209

AUTHOR Roid, Gale H.
TITLE New Technologies in the Writing of Test Items.
PUB DATE Aug 84
NOTE 19p.; Paper presented at the Annual Meeting of the American Psychological Association (Toronto, Ontario, August 1984). Tables contain small print.
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Cognitive Measurement; *College Faculty; College Instruction; Computer Assisted Testing; Criterion Referenced Tests; Diagnostic Tests; Guidelines; Higher Education; Measurement Objectives; Research Needs; *Teacher Made Tests; *Test Construction; Test Items; Test Theory

ABSTRACT

The goal of test item writing is to create measures of the true achievement dimensions desired, not extraneous abilities or related skills. To encourage university instructors to apply newer test item writing technologies to test construction for course examinations, this paper: (1) sets forth basic principles for quality item writing; (2) reviews recent research on item writing methods; (3) suggests areas for future research; and (4) provides application guidelines for instructors. Test blueprint methods match the cognitive test level to the level of instruction. Theory based methods use research theories in learning, attitude, or intelligence to dictate test construction steps. Diagnostic methods break achievement into components to identify student weaknesses. Emerging areas in item writing research include: (1) computer applications; (2) linking teaching and testing; (3) components of achievement performance; and (4) multidimensional analysis of instructionally-based tests. While an instructor's time limitations may prevent the complex analyses and field testing necessary to some of these methods, certain principles can be applied more informally and pay rich dividends. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

NEW TECHNOLOGIES IN THE WRITING OF TEST ITEMS

Gale H. Roid
Western Psychological Services
Los Angeles, California

ED255566

Paper presented as part of the invited symposium, "The technology of item writing in the teaching of psychology," at the annual meeting of the American Psychological Association, Toronto, August, 1984. Portions of this paper are to appear in Chapter 3 of the forthcoming book by R.A. Berk (Ed.) Guide to Criterion-Referenced Measurement, (2nd Ed.). Baltimore, MD: Johns Hopkins University Press, 1984. Some additional material for this paper was prepared as part of a chapter to appear in the forthcoming book by B.S. Plake and J.C. Witt (Eds.), The Future of Testing: The Second Buros-Nebraska Symposium on Measurement and Testing, Hillsdale, NJ: Erlbaum, 1985.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G.H. Roid

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 217

With the press of deadlines and hectic schedules, it is very tempting to find a convenient list of objectives or some reference materials and start writing draft items. For some subject-matter areas and for some course examinations to be used only once, an informal or objective-based method of item writing may be the only reasonable and efficient choice. However, when one wishes to design item banks for long-term usage or examinations for sophisticated diagnostic purposes, extra time spent on the rules for generating items may pay rich dividends.

Based on personal experience with college teaching, research on student study-habits, and the logic of criterion-referenced testing, there appear to be a relatively short list of important principles of quality item writing for course examinations:

1. Consistency between teaching and tests.
2. Clarity of the expected content or cognitive level of an exam as expected by students (...without a trivial revelation of the exact items of an exam)
3. Appropriateness of the cognitive level of the exam for the objectives of the course
4. Appropriateness of the item format for the purposes of the course.
5. Freedom of items from nuisance factors (Guessing, over-cueing, trivial reduction to simple computation or reading excercises)
6. Construct 'purity' of items -- do they measure the achievement in the course, or are they 'contaminated' by other abilities
7. Provision for item banking or improvement of items over time.

Tests and examinations that have an explicit blueprint or conceptual schema are more amenable to self-critique and improvement over time. In line with the first principle listed above, a conceptual schema for a university course examination would include a specification of the match between the purposes and content of the course and the purpose and content of the test items. Merrill, Reigeluth, and Faust (1979) have provided one of the better conceptual methods for analyzing the match between teaching and tests. They discuss principles for examining the acequacy of the three major elements of any course of instruction: (1) objectives or goals, (2) instructional materials or presentations, and (3) tests. Furthermore, they provide specific guidelines for assessing the consistency among these three elements. Methods of this type may be called "test blueprinting," in reference to the time-honored method of developing a matrix of content by task-level into which items are classified. As highly creative people, instructors in higher education may be more motivated by the content of the course than by the tedium of schemas, matrices, or blueprints. Unfortunately, there are many students who have not aquired the skill of anticipating the content or cognitive-level of course examinations. These students suffer most when tests contain unexpected material or tasks. One could argue that such students are inacequately "motivated," but, clearly, there needs to be a distinction between the students who poorly anticipate a test due to lack of attention or motivation and those who have exerted considerable effort but in an uninformed direction.

This paper will briefly review some of the recent work in item writing, beginning with methods that help instructors to plan tests. In

BEST COPY AVAILABLE

contrast to informal approaches, the focus of this paper will be on methods of item writing that have been variously described as domain-referenced (Hively, 1974), operationally-defined (Bormuth, 1970; Osburn, 1968), algorithmic (Durnin & Scandura, 1973), and forming a technology of item writing (Roid & Haladyna, 1980, 1982). These are typically contrasted with objective-based approaches and are described in typologies of achievement tests as having well-defined domains (Nitko, 1980; Wardrop, Anderson, Hively, Hastings, Anderson, and Muller, 1982).

Several scientific and practical advantages accompany the use of item-writing technologies: (1) the explicit description of item-writing methods so that instructors can self-critique and improve their tests over time, (2) increased efficiency in producing multiple test forms for use in mastery-testing (and retesting) systems, and (3) greater potential for linking tests to instructional methods and materials (Anderson, 1972; Berk, 1980; Bormuth, 1970; Millman, 1974; Shoemaker, 1975).

Further motivation for exploring item-writing methods comes from the exciting developments in diagnostic achievement testing, with its link to contemporary cognitive psychology (Glaser, 1981; Haertel & Calfee, 1983; Nitko, 1980), and increasing interest in the component processes of student test-item performance (e.g., Carroll & Maxwell, 1979; Brown & Burton, 1978). This work is bringing into question the assumption that achievement and aptitude tests are or must be unidimensional, and forcing more attention to the different rules or cognitive strategies that different students use to solve achievement test items of the problem-solving type. Birenbaum and Tatsuoka (1983) have shown, for example, that the "factorial structure of the [achievement test] data is highly affected by the existence of different algorithms underlying the student response patterns". (p. 23) Specific attention to the construction of test items would allow these latent cognitive processes of students to be uncovered, and new subtest scores based on these processes derived so that psychometric qualities of the test could be improved compared to conventional total scores (Birenbaum & Tatsuoka, 1983). In this way, diagnostic information about the erroneous rules used by students (problem-solving "bugs") could be uncovered and provided to instructors as feedback. Presumably, this will only be possible if achievement test items have been specifically written to provide such diagnostic information; hence, the interest in algorithmic item-writing methods.

This paper will briefly describe three major types of item-writing technologies: (1) "test blueprint" methods that provide a conceptual plan for tests, (2) theory-based methods such as the mapping-sentence method and some of the newer cognitive-component methods, and (3) diagnostic methods such as Scandura's (1977) algorithmic approach and several methods derived from experimental cognitive psychology. Examples and references for each will be given, including a review of the strengths and weaknesses of the various approaches. The paper concludes with suggestions for future research and guidelines for instructors.

Review of Item-writing Methods

Earlier reviews of the emerging technology of item writing (Berk, 1980; Engel & Martuza, 1976; Roid & Haladyna, 1980) typically emphasized a continuum of methods from the informal to the objective-based to the algorithmic. Now, with further development and perspective on the field of item

BEST COPY AVAILABLE

writing, it is possible to further differentiate the "algorithmic" end of the continuum. However, before proceeding, it should be acknowledged that the categories chosen are somewhat arbitrary. For example, some diagnostic methods are also clearly theory-based. Also, the number of different, well-described methods is small, and they remain difficult to categorize exactly because of their evolving, developmental characteristics.

In addition, it should be noted at the start that the use of any method of item writing should be accompanied by the time-honored collection of "conventional wisdom" represented in the lists of good item-writing principles. Practical guidelines for writing quality test items have been compiled by Coffman (1971), Conoley and O'Neil (1979), Ebel (1972), Gronlund (1968), Roid and Haladyna (1982), and others. These guidelines frequently give examples of multiple-choice, true-false, matching, and constructed-response items written well or poorly to point out the common flaws in test item construction. These lists of common flaws are invaluable in screening proposed item-generation algorithms, especially if the intention is to automatically generate large numbers of items.

The lists of flaws can be grouped under two major principles of good item writing: (1) the reduction of guessing factors or clues susceptible to test-wiseness, and (2) the elimination of extraneous ability factors such as reading-speed or attention-to-detail capacities. For example, in multiple-choice items, writers are instructed to make the distractors grammatically parallel and of similar length to avoid cuing the correct answer. Or, lengthy item distracters that require the reading of redundant material are seen as injecting too much unnecessary reading into the item. Basically, the task of item writing is to create measures of the true achievement dimensions desired, not extraneous abilities such as test-wiseness, or related skills such as logical thinking, reading ability, or speed of information processing.

Test Blueprint Methods

For decades, standardized achievement tests and tests for college admissions have been designed using a blueprint method that contrasts the content of the items with the cognitive level of the tasks inherent in the items. Practical application of this method at a simple level can be as easy as drawing a matrix of lines on a page and labeling the columns with cognitive levels such as "facts, concepts, principles, applications, and evaluations", and the rows with subject-matter content labels. Another practical method is to develop an item bank, either on a microcomputer or on inexpensive card files where the content and cognitive level of the item(s) is recorded along with item statistics and other information about the quality of the items.

Keller and Reigeluth (1984), in this symposium, will discuss one of the prominent methods of blueprinting tests that strikes at the heart of the problem of making the cognitive-level of tests match the level of instruction. They discuss the useful distinction first made by Merrill, et al. (1979) of the difference between "remember" "use" and "find" which are performance dimensions of the cognitive processes underlying test-item performance.

Whether one uses a schema such as Reigeluth's, or Bloom's, or typologies such as Miller, Williams, and Haladyna (1978), the major

PY AVAILABLE

point is that some attention is given to classifying and counting the types of items included in a course examination. Without an accounting, there is no benchmark for improving the cognitive-requirements of a test, and no convenient way to assess the representativeness or content/task fairness of a test in relation to instruction. The best of blueprints are actually 4-dimensional:

1. The subject-matter content areas (e.g., in a psychological assessment course these might include intelligence, personality, and interest assessment)
2. The cognitive-content of items (e.g., fact, concept, principle, or procedure)
3. The performance level of the items (e.g., are students required to remember the facts, principles, etc., or apply them, or synthesize several of them?)
4. Closely linked to number 3, but useful as a specific dimension to list -- the cognitive requirements of the format of the item (is it constructed or selected response, essay or multiple choice).

Within the constraints of practicality (e.g., efficiency and costs require that tests be machine scored), attention to these design aspects of test blueprinting, even if done informally without lengthy written records, should provide improved consistency between teaching and testing.

Theory-based Methods

A category of item-writing methods includes those based on a theoretic or research foundation that dictates the steps used to construct a set of items. In nearly ever case, these methods are an outgrowth of research in areas outside of achievement testing, such as the study of attitudes, conceptual learning, or the nature of intelligence. These methods remain interesting and of potentially great value to instructors.

Mapping sentences. Similar to the way in which item forms (Hively, 1974) are created, mapping sentences are written with variable elements (called facets) that define variations in wording that could be used to create a large number of parallel sentences. However, the major difference in the two approaches is that mapping-sentence methods were intended to be verified by empirical research (typically using some form of cluster analysis or smallest-space analysis) that documents the relationship between the theoretical structure of the items and actual data. In practice, the mapping-sentence method has been applied informally without the verification step.

Structural facet theory (Guttman, 1959; Foa, 1965) and its subsequent application to scale development and item writing (Guttman & Schlessinger, 1967; Guttman, 1970; Runkel & McGrath, 1972) forms the basis of a method which has become identified with one of its major features -- the mapping sentence. The method of mapping sentences has a practical application immediately if one focuses only on the mapping sentence itself,

BEST COPY AVAILABLE

without doing the computerized data analyses (e.g., Lingoes, 1977). Table 1 shows a possible mapping sentence for a topic in American literature, the "greatness of Mark Twain's Huckleberry Finn." Although this example has not been verified empirically on data from literary critics, it is clear that breaking the topic into a general sentence with several facets is very helpful in a practical understanding of what could be tested. If one wishes to have an empirically validated set of questions, a study would need to be conducted of representative

---

Insert Table 1 about here

---

questions generated on the topic of Huckleberry Finn that were completed by different groups such as (1) expert critics in American literature, (2) graduate students in literature, (3) other students and novices, to suggest but a few. Such a study might reveal that certain dimensions underlying the items define the acknowledged criteria by which experts judge the novel, and that more items can be written to tap these dimensions more thoroughly. To date, it is not clear that anyone has attempted such an empirical extension of facet theory into the realm of achievement testing, but the potential of such a technique is intriguing.

For further examples and guidelines for developing mapping sentences for examinations or tests, the reader is referred to the articles by Berk (1978) and Guttman and Schlesinger (1967), Chapter 8 in Roid and Haladyna (1982), and the general reference work of Runkel and McGrath (1972).

Testing concepts, rules, and principles. One of the best examples of systematic item writing based on theories of learning and teaching, is the work of Tiemann and Markle (1983). By integrating the considerable research on concept and rule learning, Tiemann and Markle developed a training package (Markle & Tiemann, 1970) that demonstrated how to analyze concepts in a variety of subject-matter areas from Art to Zoology. The basic theory of concept learning implemented by the technique included 3 ideas: (1) that students must learn to discriminate between examples and non-examples of a concept, (2) that students must generalize from teaching examples to a broader set of examples, and (3) that testing the understanding of concepts must include both examples and non-examples different from those used in teaching the concepts.

In the Tiemann and Markle system, concepts are defined as classes of entities, the members of which share some properties in common. Some concepts and their subject-matter areas are soliloquy (Literature) synonym (grammar), mercantilism (history), element (chemistry), and reliability estimate (measurement). For each concept, the critical and variable attributes of examples and non-examples can be specified. For instance, in the concept "insect", one of the critical attributes is that it must be an invertebrate with six jointed legs, but the size of the legs or the presence of wings are attributes which can vary across both examples and non-examples.

Tiemann, Kroeker, and Markle (1977) showed that a set of related concepts can be coordinated so that examples used to teach one concept may be used as non-examples for another concept. They examined the related

BEST COPY AVAILABLE

concepts of positive and negative reinforcement, punishment and extinction, and they were able to devise a criterion-referenced system of questions. The domain for the items was specified by the critical and variable attributes of the examples and non-examples used in teaching and testing. Again, examples used in teaching were excluded from the tests so that true comprehension could be assessed (Anderson, 1972).

Recently, Tiemann and Markle (1983) have extended the method to include the analysis of rules and principles, which can be viewed as statements of the relationship among concepts. It is often easier to identify rules or principles in instructional materials or teaching exercises than it is to begin by finding the more elemental concepts. For example, in the realm of writing and copy editing, a general rule for using compound words is, "When a temporary compound is used as an adjective before a noun, it is often hyphenated to avoid misleading the reader." Examples of compound words from this paper would include "item-writing method" and "subject-matter expert". A thorough study of the rule reveals that a number of fundamental concepts in language arts and grammar are embedded in it -- "temporary compound," "adjective," and "noun." By providing a comprehensive list of examples and non-examples of each concept, it is possible to construct a huge domain of teaching and testing items for assessing understanding of the rule.

Learning to use the Tiemann and Markle approach takes study time and some feedback from colleagues or subject-matter experts. An excellent place to begin is the training workbook (Tiemann & Markle, 1983) which includes many practical exercises.

Diagnostic Methods

Algorithmic methods. One of the most widely referenced methods of item writing was proposed by Scandura (1970, 1977). Durnin and Scandura (1973) described the method as an algorithmic approach, with more potential for diagnosing the problem-solving errors of students than the method of item forms. The choice of the term "algorithm" comes from the fact that problem-solving behavior, particularly in mathematics and quantitative areas of natural and social sciences, can be described by a flow-chart, similar to a computer algorithm. If the algorithm is an accurate representation of the real procedures that students use in solving problems, then the algorithm can be used to simulate the "paths" students can take through the various steps in problem solution. Also, the steps in the algorithm, once listed and described, can be used to develop a universe of possible problems, and to track student errors and incorrect choices.

Table 2 presents a sample problem in arithmetic, analyzed from an algorithmic viewpoint. Two algorithms are listed. The first was adapted from Durnin and Scandura (1973), and is a systematic method of summing the first two addends and then adding each successive number. The second, informally created for this table, is a contrasting algorithm that sums pairs of addends, and includes an attempt to increase the speed of problem solution through the pairing of those numbers that sum to 10. The two contrasting algorithms are presented in order to emphasize that even for a simple problem, there are multiple strategies or rules that students can use to solve a problem. By anticipating and specifying the steps in

such rules before a test is developed, one can design rather sophisticated diagnostic information into the items. Diagnostic information may be derived from problems that exaggerate certain errors that students are known to make. One of many possible examples might be to overload a problem such as presented in Table 2 with pairs of numbers that have sums greater than 10, to test the accuracy of the student's "carrying" into the tens position. Another example would be to include a set of items that have many pairs of addends that sum to exactly 10, to test for the presence of the efficiency strategy mentioned in Algorithm 2 in Table 2, through the recording of response time. (For further information about the method the reader is referred to Durnin and Scandura (1973), Scandura (1970, 1977), and Scandura and Durnin (1978).

Insert Table 2 about here

Other diagnostic methods. An increasing number of studies are appearing in the measurement literature concerning the design of achievement tests that are measuring separate components or steps in the performance of students on test items of the problem-solving type. By providing separate subtests or scores for each component or step, tests of problem solving can become more diagnostic of student errors. Because no formal names for such methods has been coined, nor detailed examples of sample items provided in the literature, they will be described in rather general terms.

Brown and Burton (1978) used information processing models of student performance on addition and subtraction problems, as a method of simulating the problem-solving operations commonly used by elementary students. The study used a concept from computer programming to describe errors embedded in student problem-solving strategies. These errors are called "bugs," and any programmer is familiar with the process of "debugging" a computer program. By documenting the common bugs that students acquire, anticipated wrong answers to questions can be recorded and used to identify the presence of particular errors. Remedial instruction in mathematics or statistics can be seen as a process of "debugging" the incorrect algorithms used by students in solving calculation problems. Also, the work of Brown and Burton, Scandura, and others highlights the importance of knowing the algorithms initially taught to a student, so that diagnostic tests can be linked to the actual algorithms used in learning.

Birenbaum and Tatsuoka (1983) have devised methods for scoring arithmetic tests based on the underlying algorithms used by students to solve each item. They were successful in eliminating the variability in test scores due to "The fact that students can sometimes get right answers by following the wrong rule". Patterns of student responses to signed-number problems, for example, were intensely analyzed, and students were interviewed to find the incorrect algorithms being used to solve these problems. For instance, some students converted subtraction problems into addition problems, ignoring the signs of each number. New scoring methods gave a score of "1" only when a correct answer was achieved by using an error-free algorithm. In other words, students were not given credit for a

BEST COPY AVAILABLE

"correct" answer obtained by an erroneous strategy. Results showed that the internal consistency was greater (.93 versus .84) for the new scoring methods, and the factor structure of the rescored items was greatly simplified.

In order for the Birenbaum and Tatsuoka (1983) scoring method to be used, presumably a sufficient number of items needs to be written to uncover the majority of common algorithms used by students in solving a set of problems. Therefore, their work has direct implications for the technology of item writing. Also, the approach is a further example of a "bootstrapping" method of item writing that could be very valuable. After an initial field testing of a group of items, the dimensionality underlying the test items is examined using cluster analysis, multi-dimensional scaling, or factor analysis, or through attempts to fit multicomponent latent-trait models (e.g., Reckase, 1979; Stegelmann, 1983; Whitely, 1981). Then, the acequacy of the items for measuring the latent cognitive components of student test performance is assessed. If too few items are present to assess one or more of the underlying cognitive strategies or components, the test can be expanded accordingly. A similar bootstrapping procedure was recommended in the discussion of mapping sentences presented earlier.

Suggestions for Future Research

Computer Applications

The growing use of computers in all phases of testing and test development will surely influence the directions of future research on item writing for instructionally-based tests. The implementation of item forms on microcomputers will make possible increasing research on their application to practical learning tasks. Progress in the field of linguistic transformations may require computerization in order to make the methods more immediately useful to busy educators. For example, similar to the microcomputer software that computes readability indices and corrects spelling errors as part of word-processing, a software package could identify important key words in a passage and print them for the textbook writer or reading-test developer. The Carroll, Davies, and Richman (1971) word-frequency index, which could be made accessible by such a processor, could be used to find those nouns that have a standard frequency index of 60 or less (as recommended by Roid & Finn, 1978). The sentences in which the identified words occur could then be available for printing, and, perhaps, could be partially processed for possible use as test questions. Research teams may be composed of linguists, cognitive psychologists, and educators if the technology of sentence transformation or other more complex linguistic transformations is to progress further. The integration of Bormuth's (1970) work with the contemporary work on cognitive components of reading (e.g., Curtis & Glaser, 1983) is a major challenge.

Another interesting application of microcomputers is in the development of programs that combine instruction and sophisticated diagnostic testing. Scandura (1981) has developed a system called RULE-TUTOR that implements a diagnostic, branching procedure for assessing a student's mastery of algorithms for solving problems such as those in basic mathematics. With as few as three items, one can

diagnose the student's mastery of the rules of problem solving, assuming that one has a hierarchically structured content area and an algorithmic analysis of the problems, such as was done by Durnin and Scandura (1973) and Scandura (1982). The development of computer-based diagnostic testing is certainly an area where sophisticated item-writing technologies will be needed.

Linking Teaching and Testing

Additional research is needed to guide practitioners in matching the appropriate item-writing methods to particular instructional tasks. Perhaps the concept of task structures (Baker & Herman, 1983) will help in this regard, but also, the practical guidelines such as those currently used by the U.S. Navy (Ellis & Wulfeck, 1983) need extension into the broader areas of school learning. Although the Navy method begins with performance objectives, rather than domain specifications, it is extremely comprehensive in including most of the item-writing methods discussed in this paper, and it provides a means of matching task and content characteristics to test item formats in the realm of skills training. It uses the Instructional Quality Inventory system (Montague, Ellis, & Wulfeck, 1983; Roid & Haladyna, 1982), which categorizes all objectives and test items into a grid composed of tasks (Remember, Use-Unaided, Use-Aided) by contents (Facts, Concepts, Procedures, Rules, Principles). Other typologies (e.g., Miller, Williams & Haladyna, 1978), methods of analysis (e.g., Merrill, Reigeluth, & Faust, 1979), and systems (e.g., the PLANA system of Schott, Neeb, & Wieberg, 1982) may be useful schemas for matching item types with instructional requirements.

Components of Achievement Performance

As Guttman (1970) insightfully observed, there have been many theories of test scores, but few theories of content structure and specification. It is ironic, therefore, that the growing recognition of the usefulness of item-response theories such as the Rasch model (Rasch, 1980) and the three-parameter model (Lord, 1980) has also highlighted the unidimensionality assumption of achievement tests. With increased attention to item-response theory has come rather immediate concern for multicomponent and multidimensional models (Reckase, 1979; Whitely, 1981). Hopefully, the excitement over the statistical aspects of multidimensional test analysis will not distract researchers from the conceptual and research tasks ahead in specifying the components of achievement performance. The methods proposed by Scandura (1983) for analyzing content areas, especially basic mathematics, are an excellent example of new analytical methods that attempt to uncover prototypic strategies or rules used to solve problems such as those found on many achievement tests.

The complex challenge ahead will be to further extend the important work begun on decomposing test item performance into its cognitive or information-processing components. Stimulated by work on intelligence (Sternberg, 1977, 1979, 1981) and aptitude testing (e.g., Whitely, 1977, 1981), researchers can begin to search for the component processes underlying achievement test performance, with the recognition that new test

COPY AVAILABLE

scoring and statistical models (e.g., Birenbaum & Tatsuoka, 1983; Stegelmann, 1983) will be awaiting the computer entry of real data.

Multidimensional Analysis of Instructionally-Based Tests

A final area of new research will undoubtedly be the study of the multidimensional nature of achievement test data. Just as factor analysis has often motivated psychologists to add or subtract certain items from personality and ability tests because of a desire to measure factors, perhaps educational test developers will begin to create test items to measure the underlying dimensions of new multidimensional, multicomponent achievement tests. The upsurge in interest in item-response theories, has brought an increased interest in the determination of test dimensionality (Green, Lissitz, & Mulaik, 1977; McDonald, 1981; Smith, 1980). New methods of factor analysis that recognize the problem of factoring dichotomous items may also stir interest in this much neglected area.

Gorsuch and Yagel (1982) recommended two types of analyses: item-parcel procedures and hierarchical factor analysis. Cattell (1956, 1974) has used a technique called item parceling for decades, but its usage has somehow escaped the attention of educational test developers. Three types of item parceling have been used: informal, radial, and suppressor parceling. Each will be briefly described here. Informal item parcels are mini-scales composed of 3 to 7 items judged to be similar in content, which are then scored, and the parcel scores are subjected to a factor analysis. Radial parcels are defined by an initial factor analysis that extracts a large number of factors. Then, an examination of the pattern of factor loadings is used to form parcels of similar items. In suppressor parceling, each grouping of items is composed of three or more items that differ systematically along some "nuisance" dimension that needs to be suppressed in order to study the "true" factor structure. For example, items which are high, medium, or low on guessing or test-wiseness influence (as determined by some empirical index of these influences) would be placed in each parcel. The resulting factor analysis of the parcel scores would have neutralized the influence of the nuisance variable(s).

Huba (1984) has developed an item analysis program that routinely assesses the dimensionality of dichotomous items using alternate measures of interitem association (in contrast to Pearson product-moment correlations): (1) tetrachoric correlations, which assume an underlying normal distribution behind the dichotomous response, and (2) Yule-Q statistics for 2-by-2 tables which correct for the occurance of "zero cells" (e.g., where students may never score incorrect on both items). Then, factor analyses on inter-item matrices of these measures are calculated. The size and pattern of loadings on the first, unrotated principal component extracted from such matrices are indicative of the quality of unidimensionality of the test. Also, items which do not have high loadings on the first component may have extraneous, nuisance, or qualitatively different subfactors within them that are undesirable if one is to compute a single "total score." Such routine analyses of dimensionality could easily be incorporated into item analysis programs offered by college or university examination-scoring services (with less than 1000 lines of FORTRAN coding).

In a series of important studies, Muthen (1978, 1981) and Muthen

and Christofferson (1981) have developed a factor-analytic model specifically for dichotomous items. A generalized computer program, which performs other functions of structural equation modeling, is also under development and is efficient in analyzing 20 to 25 variables at a time (Muthen, 1982). Research on the dimensionality of groups of achievement test items would certainly be an important application of these new methods.

As mentioned earlier in this paper, there are two purposes for exploring the dimensionality of course examinations or achievement tests. First, the empirical determination of the dimensions underlying test-item response data may suggest ways to write additional items having identified dimensional qualities. Secondly, the construct validity of the resulting examination or test is at stake (see Hambleton, 1980), and the ultimate purpose of writing each item on the test should be to contribute to the validity of the measure.

Guidelines for Instructors

The limited time available to instructors for the development of tests may not permit the inclusion of the complex analyses and field testing that would accompany the development of the theory-based or diagnostic methods referenced in this paper. By being aware of the need for such methods, however, the instructor can be sensitive to the search for instructional materials and tests that are diagnostic of the cognitive components used by students to solve quantitative problems, for example, in psychology courses.

At an informal level, there is no reason why certain principles of item writing cannot be applied, without the definitive application of all the steps required by one of the operational methods. Several examples come to mind. First, test blueprints continue to be a useful tool for suggesting possible test questions, even if a blueprint is not precisely specified or used to create full universes of items. Taking statements of objectives or goal statements and turning them into rows of a blueprint is often feasible as a quick planning step prior to item writing. Also, the method of mapping sentences, excluding data analysis on each facet, is equally useful in adapting statements of objectives, or themes found in textbooks. For instance, a physiological psychology text may describe each of five regions of the brain, with details given on four types of animal experiments conducted to assess the effects of lesions or stimulation to each brain location. This should rather immediately suggest a mapping sentence with two facets with 5 and 4 elements, respectively, resulting in 20 possible questions.

In the case of large-enrollment courses, resources may allow for much more elaborate test-item development than is feasible for the individual instructor. In this arena, the new extension of Tiemann and Markle's (1983) concept-based approach to rule and principle learning would be highly recommended. It seems that there is hardly a subject-matter area that is not filled with principles and rules, the bases of which are important concepts that have not yet been systematically analyzed.

As the methodology of diagnostic methods develops further, it will be exciting to see large-course exams emerge that measure the cognitive components embedded in student performance in, e.g., quantitative

BEST COPY AVAILABLE

psychology. Furthermore, the application of new scoring procedures, such as those advocated by Birenbaum and Tatsuoka (1983), would allow the documentation of the frequency (e.g., in sections of a course) of common erroneous strategies for solving quanititative problems. At the same time, the possibility of such new scores should encourage the development of quanitative subtests sensitive to the common errors made by students.

Finally, the informal application of diagnostic methods of item writing would surely pay rich dividends in teaching, especially in areas of remedial instruction. Without having to construct a full algorithm with all its flowchart steps, the instructor can note the common kinds of strategies a student uses to solve a quantitative psychology problem, and then insure that a few test items are included that would show the typical errors created by students who use an incorrect method to solve the problem. Surely, instructors have been doing this for years, but it has never been formally labeled as an algorithmic method of test construction. Perhaps a deeper appreciation that students do in fact have unique cognitive strategies developed in their own ways to solve problems will do much to stimulate attention to diagnostic testing.

References

Anderson, R.C. How to construct achievement tests to assess comprehension. Review of Educational Research, 1972, 42, 145-170.

Baker, E.L., & Herman, J.L. Task structure design: Beyond linkage. Journal of Educational Measurement, 1983, 24, 149-164.

Berk, R.A. The application of structural facet theory to achievement test construction. Educational Research Quarterly, 1978, 3, 62-72.

Berk, R.A. A comparison of six content domain specification strategies for criterion-referenced tests. Educational Technology, 1980, 20(9), 49-52.

Birenbaum, M., & Tatsuoka, K. On the dimensionality of achievement test data. Journal of Educational Measurement, 1982, 19, 259-266.

Birenbaum, M., & Tatsuoka, K. The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. Journal of Educational Measurement, 1983, 20, 17-26.

Bormuth, J.R. Comparable cloze and multiple-choice comprehension test scores. Journal of Reading Behavior, 1967, 10, 291-299.

Bormuth, J.R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Brown, J.S., & Burton, R.R. Diagnostic models for procedural bugs in basic mathematics skills. Cognitive Science, 1978, 2, 155-192.

Carroll, J.B., Davies, P., & Richman, B. Word frequency book. Boston: Houghton-Mifflin, 1971.

Carroll, J.B., & Maxwell, S. Individual differences in ability. Annual Review of Psychology, 1979, 603-640.

Cattell, R.B. Validation and intensification of the Sixteen Personality Factor Questionnaire. Journal of Clinical Psychology, 1956, 12, 205-214.

Cattell, R.B. Radial parcel factoring versus item factoring in defining personality structure in questionnaires: Theory and experimental

ST COPY AVAILABLE

checks. Australian Journal of Psychology, 1974, 26, 103-119.
Clarke, A.C. The challenge of the sea. New York: Holt, Rinehart and Winston, 1960.
Coffman, W.E. Essay examinations. In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 1971, Pp. 271-302.
Conoley, J.C., & O'Neil, H.F., Jr. A primer for developing test items. In H. F. O'Neil, Jr. (Ed.), Procedures for instructional systems development. New York: Academic Press, 1979, Pp. 95-127.
Curtis, M.E., & Glaser, R. Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 1983, 20, 133-148.
Durnin, J., & Scandura, J.M. An algorithmic approach to assessing behavior potential: Comparison with item forms and hierarchical technologies. Journal of Educational Psychology, 1973, 65, 262-272.
Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.
Ellis, J.A., & Wulfeck, W.H. Handbook for testing in Navy schools. San Diego, CA: Navy Personnel Research and Development Center, 1983.
Engel, J.D., & Martuza, V.R. A systematic approach to the construction of domain-referenced multiple-choice test items. Paper presented at the annual meeting of the American Psychological Association, Washington, DC, September 1976.
Fox, W.S. Three kinds of behavioral changes. Psychological Bulletin, 1968, 70, 460-473.
Glaser, R. A research agenda for cognitive psychology and psychometrics. American Psychologist, 1981, 36, 923-936.
Gorsuch, R.L. Factor Analysis (2nd ed.). Hillsdale, NJ: Erlbaum, 1983.
Gorsuch, R.L., & Yagel, J.C. Exploratory item factor analysis. Unpublished paper, Fuller Graduate School of Psychology, Pasadena, California, 1982.
Gorsuch, R.L., & Dreger, R.M. "Big jiffy": A more sophisticated factor analysis and rotation program. Educational and Psychological Measurement, 1979, 39, 209-214.
Green, S.B., Lissitz, R.W., & Mulaik, S.A. Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 1977, 37, 827-838.
Gronlund, N.E. Constructing achievement tests. Englewood Cliffs, NJ: Prentice-Hall, 1968.
Guttman, L. A structural theory for intergroup beliefs and actions. American Sociological Review, 1959, 24, 318 328.
Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 1968, 33, 469-506.
Guttman, L. Integration of test design and analysis. In Proceedings of the 1969 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1970.
Guttman, L., & Schlesinger, I.M. Systematic construction of distractors for ability and achievement testing. Educational and Psychological Measurement, 1967, 27, 569-580.

BEST COPY AVAILABLE

Haertel, E., & Calfee, R. School achievement: Thinking about what to test. Journal of Educational Measurement, 1983, 20, 119-132.

Haladyna, T.M., & Roid, G.H. The role of instructional sensitivity in the empirical review of criterion-referenced test items. Journal of Educational Measurement, 1981, 18, 39-53.

Haladyna, T.M., & Roid, G.H. A comparison of two approaches to criterion-referenced test construction. Journal of Educational Measurement, 1983, 20, 271-282.

Hambleton, R.K. Test score validity and standard-setting methods. In R.A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1980, Pp. 80-123.

Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14(6), 5-10.

Hively, W., Patterson, H.L., & Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Hubs, G.J. Program for item and scale analysis. Unpublished computer program, Western Psychological Services, Los Angeles, California, 1984.

Kingsbury, G.G., & Weiss, D.J. Relationships among achievement level estimates from three characteristic curve scoring methods (Research Report 79-3). Minneapolis: Department of Psychology, University of Minnesota, April 1979.

Lingoes, J.C. (Ed.) Geometric representations of relational data. Ann Arbor, MI: Mathesis Press, 1977.

Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Markle, S.M., & Tiemann, P.W. Really understanding concepts. Champaign, IL: Stipes, 1970.

McDonald, R.P. The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 1981, 34, 100-117.

Merrill, M.D., Reigeluth, C.M., & Faust, G.W. The instructional quality profile: A curriculum evaluation and design tool. In H.F. O'Neil, Jr. (Ed.), Procedures for instructional systems development. New York: Academic Press, 1979, Pp. 165-204.

Miller, H.G., Williams, R.G., & Haladyna, T.M. Beyond facts: Objective ways to measure thinking. Englewood Cliffs, NJ: Educational Technology Publications, 1978.

Millman, J. Criterion-referenced measurement. In W.J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974, Pp. 311-397.

Montague, W.E., Ellis, J.A., & Wulfeck, W.H. Instructional quality inventory: A formative evaluation tool for instructional development. Performance and Instruction, 1983, 22(5), 11-14.

Muthen, B. Contributions to factor analysis of dichotomous variables. Psychometrika, 1978, 43, 551-560.

Muthen, B. Factor analysis of dichotomous variables: American attitudes toward abortion. In E. Borgatta & D.J. Jackson (Eds.), Factor analysis and measurment in sociological research: A multidimensional perspective. San Francisco: Sage Publications, 1981.

Muthen, B., & Christoffersson, A. Simultaneous factor analysis of

BEST COPY AVAILABLE

dichotomous variables in several groups. Psychometrika, 1981, 46, 407-419.
Nitko, A.J. Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 1980, 50, 461-485.
Osburn, H.G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.
Reckase, M.D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.
Roid, G.H., & Finn, P.J. Algorithms for developing tests questions from sentences in instructional materials (NPRDC Technical Report No. 78-23). San Diego, CA: Navy Personnel Research and Development Center, 1978.
Roid, G.H., & Haladyna, T.M. The emergence of an item-writing technology. Review of Educational Research, 1980, 50, 293-314.
Roid, G.H., & Haladyna, T.M. A technology for test-item writing. New York: Academic Press, 1982.
Roid, G.H., Haladyna, T.M., Shaughnessy, J., & Finn, P.J. Item writing for domain-based tests of prose learning. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.
Roid, G.H., & Wendler, C.L.W. Item bias detection and item-writing technology. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.
Runkel, P.J., & McGrath, J.E. Research on human behavior: A systematic guide to method. New York: Holt, Rinehart and Winston, 1972.
Scandura, J.M. Role of rules in behavior: Toward an operational definition of what (rule) is learned. Psychological Review, 1970, 77, 516-533.
Scandura, J.M. Problem solving: A structural process approach with educational implications. New York: Academic Press, 1977.
Scandura, J.M. Microcomputer-based system for authoring, diagnosis, and instruction in algorthmic content. Educational Technology, 1981, 21(1), 13-19.
Scandura, J.M. Structural analysis (cognitive tasks): A method for analyzing content. Part I: Background and experimental findings. Journal of Structural Learning, 1982, 7, 101-114.
Scandura, J.M., & Durnin, J. Assessing behavior potential: Adequacy of basic theoretical assumptions. Journal of Structural Learning, 1978, 6, 3-47.
Schott, F., Neeb, K.E., & Wieberg, H.J.W. A general procedure for the construction of content-valid items for goal-oriented teaching and testing. Studies in Educational Evaluation, in press.
Shoemaker, D.M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.
Smith, J.K. On the examination of test unidimensionality. Educational and Psychological Measurement, 1980, 40, 885-889.
Stegelmann, W. Expanding the Rasch model to a general model having more than one dimension. Psychometrika, 1983, 48, 259-267.
Sternberg, R.J. Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum, 1977.

BEST COPY AVAILABLE

Sternberg, R.J. The nature of mental abilities. American Psychologist, 1979, 34, 214-230.

Sternberg, R.J. Testing and cognitive psychology. American Psychologist, 1981, 36, 1181-1189.

Sternberg, R.J., & Wagner, R.K. Measuring "intelligence in the real world." Paper presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

Tiemann, P.W., & Markle, S.M. Analyzing instructional content: A guide to instruction and evaluation (2nd ed.). Champaign, IL: Stipes, 1983.

Tiemann, P.W., Kroeker, L.P., & Markle, S.M. Teaching verbally-mediated coordinate concepts in an on-going college course. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Trilling, L. Introduction. In M. Twain, The adventures of Huckleberry Finn. New York: Holt, Rinehart and Winston, 1948.

Wardrop, J.L., Anderson, T.H., Hively, W., Hastings, C.N., Anderson, R.I., & Muller, K.E. A framework for analyzing the inference structure of educational achievement tests. Journal of Educational Measurement, 1982, 19, 1-18.

Whitely, S.E. Some information-processing components of intelligence test items. Applied Psychological Measurement, 1977, 1, 465-476.

Whitely, S.E. Measuring aptitude processes with multicomponent latent trait models. Journal of Educational Measurement, 1981, 18, 67-83.

BEST COPY AVAILABLE

Table 1

Example of a Mapping Sentence in American Literature

Mapping Sentence:	"The greatness of *Huckleberry Finn*'s (A) is in Mark Twain's (B) the (C)."
Face Elements:	(A) topic or theme character development form style social commentary
	(B) attention to shades of departure from
	(C) relationship of Huck and Jim. moral struggle of Huck. Divinity of the River. use of realistic American dialects. picaresque novel form. poor grandiose "literary" style.
Possible Short-Answer Essay Question Form:	"Briefly describe three episodes from the novel *Huckleberry Finn* that are examples of its greatness of (A) in Mark Twain's (B) the (C)."

Source: Derived from a reading of Trilling (1948).

Table 2

Analysis of Algorithms for an Addition Problem

Problem:	Addition of multiple, single-digit addends (Four or more addends)
Example:	3 + 7 + 4 + 9 + 5 = ?
Algorithm 1:	1. Sum the first two addends. 2. If the sum is a two-digit number go to Step 4, otherwise go to Step 3. 3. Add the units digit of the sum to the next addend and go back to Step 2. 4. Add 1 to whatever is in the tens place and go back to Step 3 if there is another addend, otherwise to Step 5. 5. Write down the sum and stop.
Algorithm 2:	1. Add successive pairs of two addends and write out these temporary sums to the side, including any remaining single addend. If any pairs add to ten, even if unconfigural, add them together first. 2. Add the "ones digits" of the temporary sums, incrementing the tens place by 1 as many times as required. 3. If the number of temporary sums is large (e.g., greater than 5), repeat Step 1 for these sums. 4. Write down the final sum.

equals 10 or more

BEST COPY AVAILABLE

From Berk, R.A. (Ed.) A guide to criterion-referenced test construction. Baltimore, MD: Johns Hopkins University Press, in press, 1984, by permission.